New Metrics for Quantifying Underrepresentation in Biomedicine and its Trends

Huifeng Yu: University at Albany, SUNY
Gerald Marschke: University at Albany, SUNY
Allison Nunez: University at Albany, SUNY
Bruce A. Weinberg: The Ohio State University

Presentation

How has the status of women, underrepresented racial and ethnic groups, and early career investigators evolved in the sciences? What factors can help understand and potentially accelerate gains? These questions have troubled policy makers and have been the subject of high-level commissions and initiatives for years because they bear on the efficiency, sustainability, representativeness, and equity of the research enterprise. Unfortunately, our understanding of them has been “stymied” by a lack of data and metrics. These problems are particularly acute because some groups are sufficiently underrepresented that our ability to study them reliably at anything other than the population scale is quite limited.

We contribute to our understanding of underrepresentation by systematically investigating author order as a metric for researcher standing. Author order is available at the population scale and, in the biomedical sciences, indicates the role that researchers play on papers, with first authors typically having primary responsibility for research and last authors typically leading research teams as principal investigators. Used in this way, author order constitutes an indicator of authors’ professional standing.

With author order as a measure of outcomes, we study trends in underrepresentation and their drivers by applying an innovative mix of machine learning methods to develop population-scale, longitudinal data on 1.93 million U.S. authors on 2.98 million biomedical research articles over a 42-year period to study differences and trends in the standing of researchers from different backgrounds. Biomedical research is a natural environment for our analysis—it is the single largest area of scientific research in the US, one where considerable concerns have been raised, and features authorship conventions that are amenable to our analysis. We use MEDLINE data from 1967-2009 and impute author-level demographic information. In particular, we use Author-ity data (Torvik & Smalheiser, 2009) to determine author identity (disambiguation); Genni (Smith, Singh, & Torvik, 2013) gender prediction data to determine gender; MapAffil (Torvik, 2015) to allow us to focus on researchers with a U.S. affiliation; and ethnicolr’s machine learning algorithm (Laohaprapanon & Sood, 2017) to determine race/ethnicity.

Our main analysis consists of logistic regressions of author position as a function of the publication year of the article; the person’s career age, ethnicity, gender, and race; interactions between these author characteristics and time; and other characteristics of the article and/or author. Our analysis employs conditional logit models that control for article fixed effects, effectively conditioning on all aspects of the article including funding, team size, quality, and journal placement, to produce estimates of how the characteristics of each author are related to the probability that a given author is, for instance, first author.

To investigate two potential explanations for trends in underrepresentation and probe the ways in which author order can best be used to study underrepresentation, we consider changes in research team size and NIH funding status and their interactions with career age, gender, race and ethnicity. Team size affects author position in two ways. Because there can only be one first author and one last author on an article, as the number of authors increases, the share of first and last authorships declines mechanically. Beyond this, team size may affect different demographic groups differently. We also consider changes in NIH funding. NIH funding decisions have tended to favor Whites over other racial and ethnic groups (Ginther et al., 2011), which could translate into different status on NIH-funded articles.

Author order provides a valuable and underutilized measure of researcher standing. To the best of our knowledge, we are the first to study how institutional and structural factors are related to author position and may contribute to changes in author position.

References

Ginther, D. K., Schaffer W. T., Schnell J., Masimore B., Liu, F., Haak L. L., & Kington R. (2011). Race, ethnicity, and NIH research awards. *Science, 333*(6045), 1015-1019. doi: 10.1126/science.1196783

Laohaprapanon, S. & Sood, G. (2017). ethnicolr [Algorithm]. Available from https://github.com/appeler/ethnicolr

Smith, B. N., Singh M., & Torvik V. I. (2013). A search engine approach to estimating temporal changes in gender orientation of first names. *JCDL '13 Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries,* 199-208. doi: 10.1145/2467696.2467720

Torvik, V. I. & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data, 3*(3), 1-29. doi: 10.1145/1552303.1552304

Torvik, V. I. (2015). MapAffil: A bibliographic tool for mapping author affiliation strings to cities and their geocodes worldwide. *D-Lib Magazine, 21*. doi: 10.1045/november2015-torvik